CLONEID: A Framework for Longitudinal Integration of Phenotypic and Genotypic Data to Monitor and Steer Subclonal Dynamics

Thomas Veith[+2], Richard Beck[+2], Vural Tagal[+2], Tao Li[+2], Saeed Alahmari[1], Jackson Cole[2], Daniel Hannaby , John Kyei[2], Xiaoqing Yu[3], Konstantin Maksin[8], Andrew Schultz[2], HoJoon Lee[5], Issam ElNaqa[4], Steven Eschrich[3], Hanlee P. Ji[5], Aaron Diaz[6], and Noemi Andor[2]

[1]Department of Computer Science, Najran University, Najran 66462, Saudi Arabia
[2]Department of Integrated Mathematical Oncology, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[3]Department of Biostatistics and Bioinformatics, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[4]Department of Machine Learning, H Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA
[5]Department of Medicine, Stanford University, CA, USA
[6]Department of Epidemiology & Biostatistics, University of California San Francisco, CA, USA
[7]Department of Radiology & Biomedical Imaging, University of California San Francisco, CA, USA
[8]The University of Commerce and Services, Medical Department (WSHIU), Poznan, Poland

[+]Authors contributed equally to this work [*]Corresponding author: noemi.andor@gmail.com

Abstract

Understanding how genetic and phenotypic diversity emerges and evolves within cancer cell populations is a fundamental challenge in cancer biology. CLONEID is a novel framework designed to organize and analyze clone-specific measures as structured time-series data. By integrating and monitoring genotypic and phenotypic experimental data over time, CLONEID facilitates hypothesis-driven and hypothesis-generating research in cancer biology. This article outlines the development, utility, and applications of CLONEID, emphasizing its role in overcoming challenges in data reproducibility, mathematical modeling, and multi-modal data integration. A webportal to the CLONEID database is available at dev.cloneid.org.

1 Introduction

Cancer is a dynamic, phenotypically diverse disease driven by underlying genetic, epigenetic, and transcriptional variation. Within a single tumor, distinct subpopulations of cells emerge with different genotypes

and gene expression programs [1]. This *intra-tumor heterogeneity* (ITH) is evident in live-cell and clinical imaging studies: intravital microscopy and functional medical imaging have directly visualized coexisting phenotypic differences, such as variations in proliferation rates, invasive behavior, metabolic activity, and drug responsiveness among cells in the same tumor[2, 3]. These observations highlight remarkable plasticity and adaptability in cancer, as genetically similar cells can adopt divergent states in response to microenvironmental cues or therapy pressure.

Numerous computational tools have been developed to quantify ITH using data from bulk-, single-cell and spatial omics technologies [4, 5, 6, 7, 8]. Despite these advances, there is currently no unified framework to compare and track the outputs of different ITH quantification methods across time or across studies. Moreover, a significant gap remains between imaging and sequencing modalities: genomic sequencing datasets often lack crucial metadata such as lineage history, sampling time, or growth conditions, while imaging experiments rarely include direct genetic or transcriptomic readouts. This disconnect hampers efforts to connect phenotypic traits observed in living cells and tissues with the underlying molecular clones characterized by sequencing and other high-throughput assays.

We introduce CLONEID to help bridge these gaps. We begin by outlining how CLONEID's infrastructure facilitates collection, connection, and analysis of genotypic and phenotypic data from in-vitro experiments, in-vivo and clinical studies, across laboratories and timeframes. We also introduce the first database compiled using CLONEID, detailing its structure and contents. Next, we demonstrate CLONEID's utility through three use cases, categorized into two broad applications: hypothesis-driven mathematical modeling and hypothesis-generating association studies or data mining. Each use case is supported by a multi-modal dataset integrated into the CLONEID database, comprising both genotypic and phenotypic data. CLONEID is available at dev.cloneid.org as a ready-to-use platform that allows researchers to upload and analyze MRI, digital pathology, and microscopy images as phenotypic snapshots, and any output from clonal deconvolution or clustering algorithms as genotypic snapshots.

2 Methods and Framework

CLONEID is implemented as a three-tier framework consisting of: (i) a user interface for data visualization and analysis, (ii) a core functionality layer implemented in Java, and (iii) an SQL database for data storage and organization. The CLONEID database provides a centralized, structured resource that can store clone-specific data, including size, expression, single-cell DNA/RNA sequencing profiles, and lineage information.

These components are structured into two modules: the phenotype module, which handles image analysis and integration of metainformation, and the multi-omics module, which manages data on the clonal composition of a sample or biopsy as derived from high-throughput assays.

Both modules are available as ready-to-use systems, enabling researchers to seamlessly upload and integrate phenotypic with genotypic snapshots. Uploaded data are managed using a secure warehousing system with a proper access control model. Phenotypic data is preprocessed and visualized using version-controlled applications (e.g. Cellpose [10], MOOSE [11]).

2.1 Phenotype module monitors growth dynamics from routine image analysis

Mounting evidence demonstrates that each cell division can introduce mutations, enabling cells to better adapt to their environment [12, 13, 14, 15, 16, 17]. This underscores the critical need for real-time monitoring of cellular changes. Traditional cell culture management software (e.g., CultureTrax, Benchling) typically provides isolated snapshots of cell states [18, 19], restricting our ability to track the emergence, persistence, or transitions of phenotypes. Moreover, imaging data acquired in clinical, in-vivo, or in-vitro contexts usually remain disconnected from genomic and high-throughput omics analyses derived from biopsies. As a result, valuable spatial and temporal dimensions are lost, hindering comprehensive analysis of

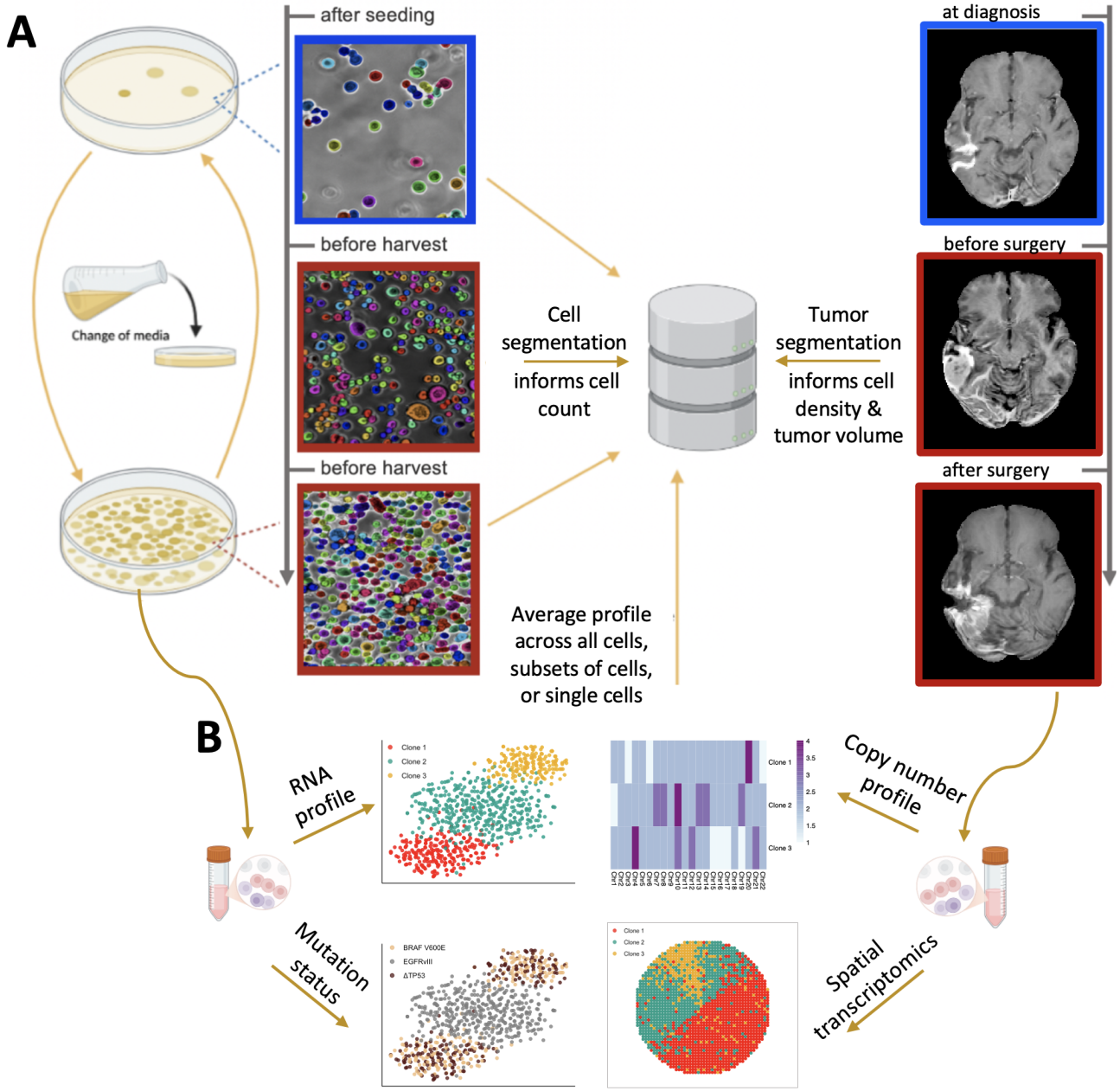

Figure 1: CLONEID workflow and central components. CLONEID is designed for multi-modal, longitudinal integration of phenotypic and genotypic data, explicitly linking genotype to phenotype across experimental or clinical timelines. (**A**) Phenotype module: each event (e.g., seedings, harvests, injections, resections) is recorded with precise timestamps linked to associated phenotypic measurements such as cell counts, media conditions, growth rates or tumor volumes. Bright-field images (*in-vitro* NUGC-4, segmented with Cellpose) and clinical radiology scans from a glioblastoma patient illustrate the real-time nature of this phenotyping. (**B**) Genotype module: Single-cell and spatial assays identify clonal populations and map their molecular states back to events recorded in the phenotype module. From left to right: (i) t-SNE embedding of scRNA-seq colored by inferred clone; (ii) clone-specific copy-number heatmap across Chr 1–22; (iii) the same embedding colored by TISCC-seq informed [9] driver mutation (BRAF V600E, $\Delta TP53$, EGFRvIII); and (iv) spatial transcriptomic slice depicting patterns of clonal expansion. Panels A and B created with BioRender.com.

clonality and evolutionary dynamics.

CLONEID's phenotype module unifies phenotypic and temporal data through an event-based capture system that is streamlined for experimental use (Figure 1A). Researchers can log in-vitro, in-vivo, or

clinical events in real time by simply naming an image and linking it to an existing entry in the database. For in-vitro studies, common events such as cell *seeding* and *harvesting* are recorded with minimal input. In-vivo procedures like cell *injections* are captured alongside mouse-specific metadata, while clinical actions such as *resection*, *diagnosis*, and *follow-up* are also logged in a consistent format.

Once an image is uploaded and linked, downstream analytic modules automatically extract and store image-derived sample characteristics—such as cell segmentation and cell counts—ensuring reproducible, longitudinal data capture. At the heart of the system is the `Passaging` table, which records every propagation or sampling event—whether an in-vitro passage or clinical biopsy—along with lineage, date, event type, and other metadata (e.g. media conditions). By linking each time-stamped entry to its sample of origin, CLONEID enables users to reconstruct detailed cell ancestries, track environmental exposure, and connect phenotypic traits to underlying genotypes across time.

Unlike conventional systems that treat samples as static identifiers, CLONEID captures the evolving context of each biological entity—without burdening the user with complex data entry.

2.2 Multi-omics module monitors changes in clonal composition

CLONEID's multi-omics module is designed to ingest outputs from widely-used clonal deconvolution or clustering tools and integrate them with rich phenotypic and temporal metadata. The core of this module is the `Perspective` table, which holds omics information at different levels of cellular resolution: (i) averaged over all profiled cells, (ii) at the level of defined clones (i.e., subpopulations), or (iii) for individual cells (e.g., from scRNA-seq or scDNA-seq). Each perspective represents a specific data view—a modular, assay-specific snapshot of a sample or subpopulation at a given time and resolution.

To capture the relationship between cell-level and clone-level data, CLONEID expects inputs at up to two levels of resolution: first, the properties of each clone (e.g., mean expression or copy number profile), and second, the member cells associated with that clone. Importantly, CLONEID stores not only the inferred clonal identities but also the input data used to generate them. For example, if clones are defined based on scRNA-seq–derived copy number states, both the copy number profiles and the underlying expression matrices are retained. This ensures that each clone's definition is traceable and reproducible, and enables users to compute or visualize additional features—such as clone-specific transcriptomic signatures—on demand.

A second table, `Identity`, integrates multiple perspectives to approximate the clonal composition of a sample. While a `Perspective` represents what we observe—a specific view from a given data source—`Identity` captures what we infer—a biologically meaningful synthesis across views that defines what a clone is. This schema supports reasoning about clonal structure across spatial and temporal samples, identification of shifts in clonal composition, and quality control for clone detection and lineage tracking.

Both `Perspective` and `Identity` tables are linked to CLONEID's phenotype module, enabling users to trace not only the phenotype of profiled cells, but also that of their ancestors and the duration each has spent in a particular environment.

2.3 Data Ingestion and Annotation Workflow

The ingestion process begins by registering each new entity into the database with a unique and persistent identifier. For cell lines, this includes experimental metadata such as origin and year of establishment. For patients, the system registers clinical metadata such as diagnosis or anatomical site. Once registered, CLONEID tracks all relevant events—such as passaging, media changes, treatments, or biopsies—along with their precise timings, ensuring longitudinal consistency across both experimental and clinical datasets.

When data acquisition events occur—such as microscopy imaging, digital pathology, or multi-omics profiling—they are registered in the system and linked directly to the specimen's position in its lineage

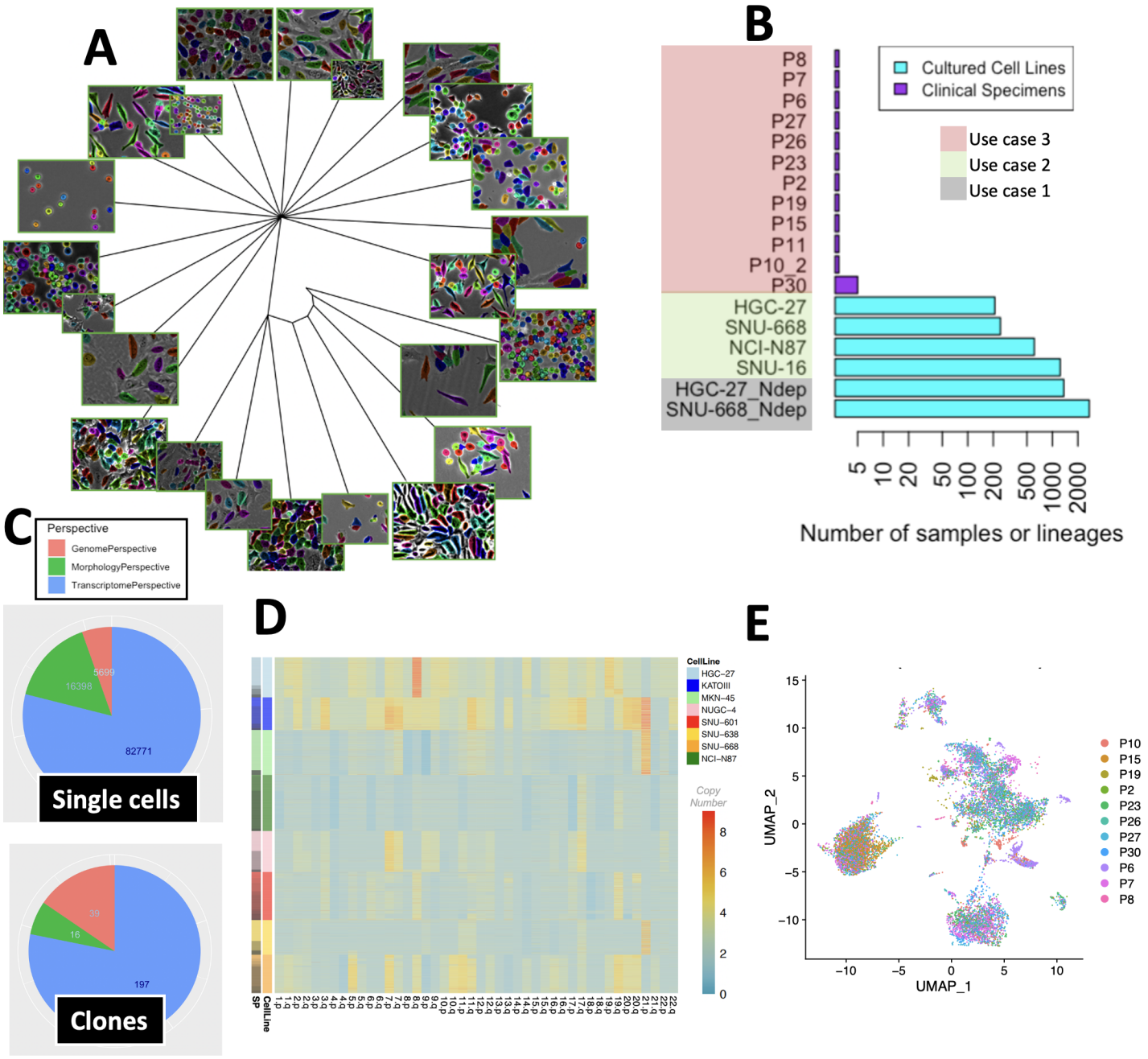

Figure 2: Database content. (A) Pedigree of the gastric cancer cell line HGC-27, with 23 lineages. Representative cells for each lineage are shown along with their segmentation masks. Pedigree trees in our CLONEID database each have between 2 - 2,726 lineages (mean=165; median=9), all accompanied by high-quality images and segentation masks. (B) Number of lineages per case represented in the Passaging table, color coded by use case and origin. (C) Number of clones (bottom) and single cells (top) represented in the Perspective table. (D) Chromosome-arm level copy number profiles across 5,779 cells from eight cell lines included in the database displayed as heatmap. (E) Gene expression profiles derived from 59,251 scRNA-sequenced glioma cells across eleven GBM patients, displayed as UMAP.

tree. This ensures that data points are contextualized within the clone's evolutionary and experimental history.

For single-cell and spatial omics data, CLONEID supports hierarchical annotation at multiple levels: (i) Subpopulation level, reflecting shared clonal identity or state-specific expression profiles and (ii) Cellular level, allowing the resolution of heterogeneity within clones or tissue regions. These annotations are stored in modular data layers (e.g., GenomePerspective, TranscriptomePerspective) and linked via unique identifiers to ensure integrative analysis across data types and time points.

CLONEID supports two modes of access: unregistered users may upload and download data freely, but their entries are automatically deleted within one hour to ensure database integrity and privacy. Users who wish to persist their entries in the database may apply for registration. Once registered, their data becomes part of the permanent record and can contribute to long-term analyses and collaborative projects.

2.4 Overview of Database Content

As of this publication, the CLONEID database contains over 5,000 lineages and specimens across 18 clinical and in-vitro cases, encompassing data types ranging from single-cell DNA and RNA sequencing to high-resolution morphological features. More than 30,000 microscopy, digital pathology, and medical images have been recorded—each embedded within large, branching pedigree trees that track cellular evolution over time (Figure 2A).

The multi-omics module hosts data from 104,896 individual cells and 252 distinct clones, organized into modular layers such as *GenomePerspective* and *TranscriptomePerspective* (Figure 2B). These layers are linked to clonal identities through the *Perspective* and *Identity* tables, enabling downstream analysis of clonal selection, tumor evolution, and genotype–phenotype relationships in full temporal and lineage context.

Users can retrieve bundled longitudinal datasets by clicking on an internal node within any pedigree tree (Fig. 2A). All descendants of that node—including associated omics profiles, image-derived phenotypes, and experimental metadata—are downloaded together, making the data instantly usable for downstream applications such as mathematical modeling, machine learning, or data mining.

3 Use Cases

We present three use cases demonstrating the utility of CLONEID. Two leverage CLONEID data to test hypotheses in vitro or in clinical settings, while the other one is a hypothesis-generating association study using in vitro data. All use cases integrate both genotypic and phenotypic data stored in the CLONEID database.

Use Case 1: Tracking Clonal Adaptation Under r/K-Selection in Gastric Cancer Cell Lines

Type of Use Case: Mechanistic, In-Vitro

We tested the impact of ecological selective pressures on a gastric cancer cell line (SNU-668). Evolution experiments were conducted by repeatedly passaging cells under low or high confluence conditions for over six months. Under r-selection, cells were seeded at low density (Fig. 3B) and allowed to grow exponentially before being re-plated, while K-selection involved high-density seeding with transfer after growth plateaued. Each lineage evolved for more than 40 passages under its respective regime. To assess phenotypic consequences of this long-term selection, we measured growth rates across density conditions (Fig. 3C). As expected, lineages evolving under r-selection exhibited significantly higher growth rates than those under K-selection, as they had more space to grow.

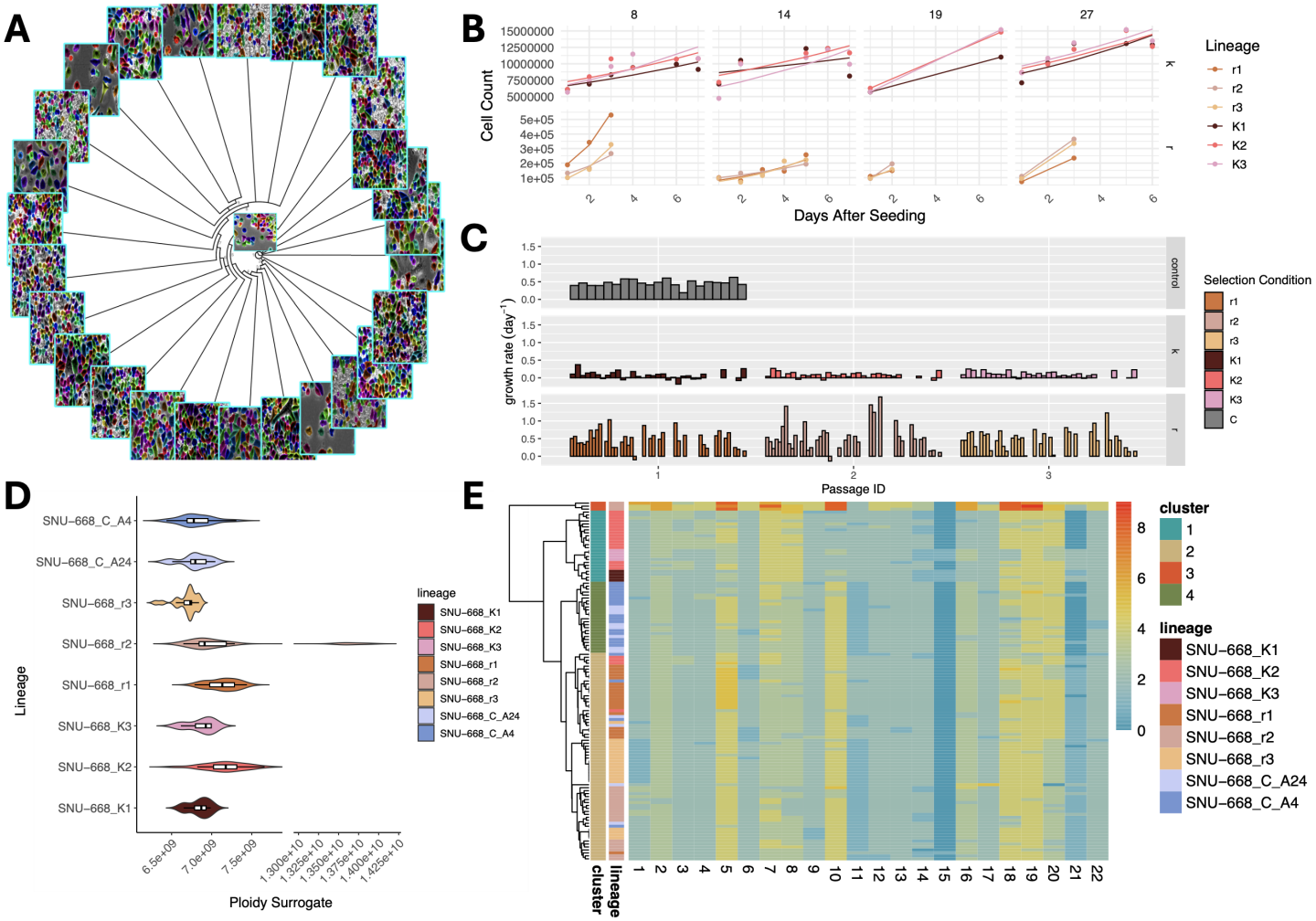

Figure 3: CLONEID reveals the impact of density-dependent selection on cancer evolution. (A) Pedigree of a subset of the K-selected SNU-668 cells, showing 9 lineages. Representative cells for each lineage are shown along with their segmentation masks. (B) Daily cell counts (datapoints) over time and fitted growth curves are shown over multiple passages (columns) for six replicates of adaptation to sparse (r-selection) or dense (K-selection) population conditions (rows and color legend). (C) Changes in growth rate (y-axis) over time (x-axis) for all replicates and passages are shown. All information displayed in (B,C) was retrieved directly by mining the CLONEID database. (D-E) Karyotype evolution can facilitate adaptation to density conditions in SNU-668. (D) Two control (blue) and six test samples from the SNU-668 cell line were karyotyped at late passages of their adaptation to their specific density conditions. Cumulative area of all segmented chromosomes was used as surrogate of ploidy (x-axis). Whole genome doubling observed in the r2 lineage. (E) Clustering cells (y-axis) according to their chromosome level copy number (x-axis) shows that cells cluster according to their density conditions (color bar).

After being cultured under these conditions for 27 and 55 passages, for K-selection and r-selection respectively, samples from each of the six test condition replicates were sent for karyotyping along with two samples from the parental lineage as controls (28 K-selected cells, 61 r-selected, and 32 control). Clustering these 121 cells by karyotype we see clusters forming by condition: most cells from the test conditions have regained a copy of chromosome 21, relative to the control lineages (Fig. 3E). In addition we note several differences between r-selected and K-selected cells, including loss of chromosomes 10 and 14 in K cells (Fig. 3E). Karyotyping revealed that r-selection also enriched for tetraploidized cells (Fig. 3D-E), indicating that whole genome doubling may be favored in early, low-competition environments. These data suggest that population density and resource availability are key features of early tumorigenesis that

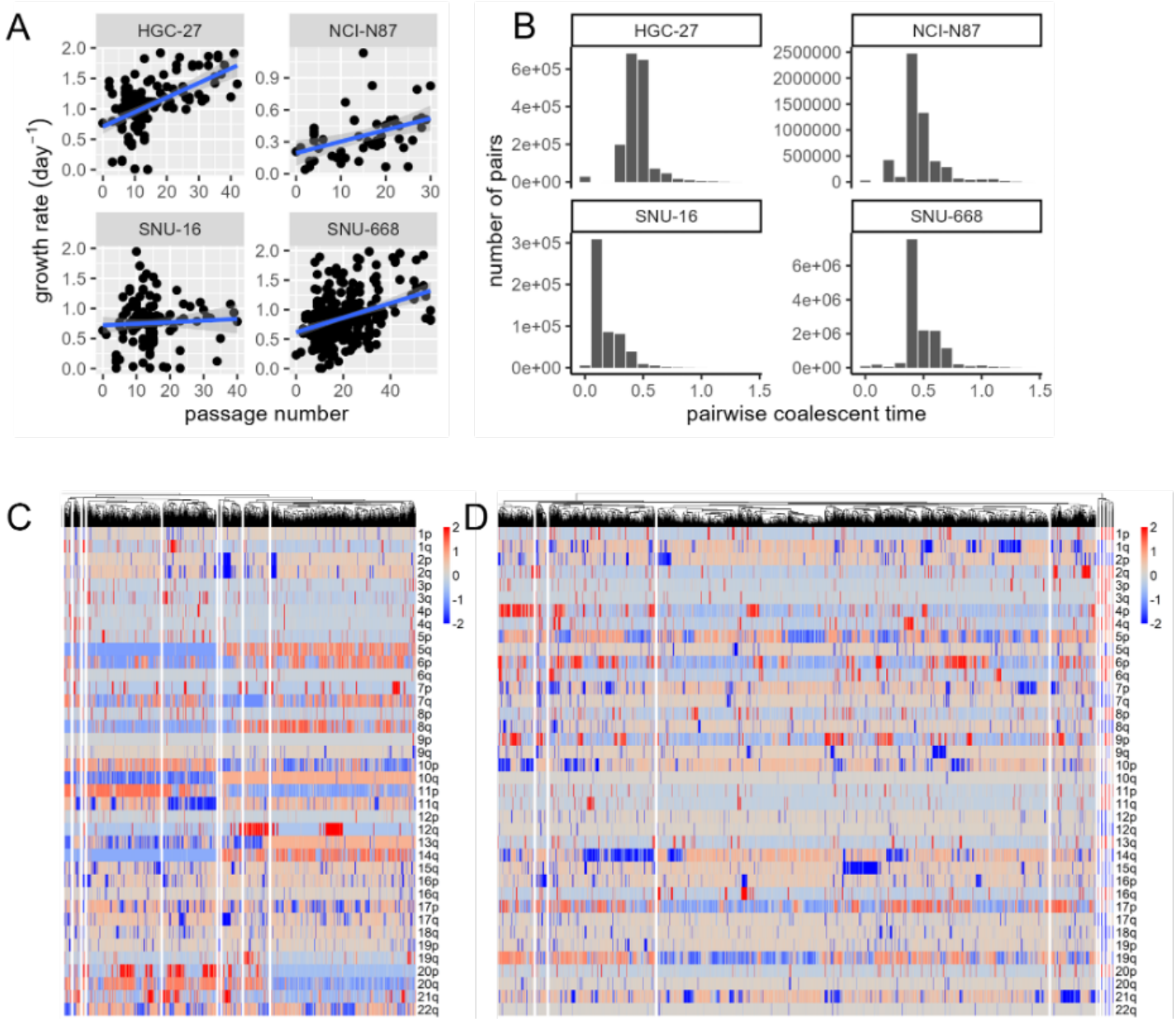

Figure 4: Predicting cellular plasticity from a snapshot of the transcriptome. (A) Passaging data were used to reconstruct lineage networks, and logistic growth models were fit to cell count trajectories from passages with at least three time points to estimate intrinsic growth rates. These growth rates were plotted against passage number to reveal temporal changes in proliferative dynamics. (C,D) CONICSmat was applied to scRNA-seq data to infer large-scale copy-number variations from normalized expression matrices. After scaling and thresholding the data, hierarchical clustering produced heatmaps with annotated cluster IDs. (B) Subsequent dendrogram analysis enabled computation of coalescent times to summarize clonal evolution. Lineages with growth rates that increase across successive passages tend to exhibit higher coalescent times in the CNV-derived dendrograms, indicating higher clonal heterogeneity may permit adaptation.

can shape the evolutionary trajectory of cancer cells.

Experiments with gastric cancer cell lines under varying population densities demonstrated the utility of CLONEID in monitoring adaptation. Data collected informed decisions about further profiling, such as karyotyping, and enabled the fitting of population growth models to observe phenotypic changes.

Use Case 2: Predicting Cellular Plasticity from a Snapshot of the Transcriptome

Type of Use Case: Data Mining/ML, In-Vitro

Four gastric cell lines in the CLONEID database (HGC-27, NCI-N87, SNU-16, and SNU-668) have evolved over 40 or more passages, representing months of continuous culture. Near-daily images of the growing cells were taken and analyzed to quantify cell numbers. A logistic growth model was fit to these cell count data to investigate changes in growth rate over time. In three of the four cell lines, significant increases in growth rate were observed, indicating adaptation to culture conditions (Figure 4).

Having observed these divergent growth trajectories, we next asked whether the underlying genetic landscape—specifically copy-number alterations—could explain the differential adaptation. To predict adaptation from transcriptomic data, copy-number-based phylogenetic trees were constructed using CONICs [20]. For each tree, pairwise coalescence times were computed for sequenced cells. For SNU-16, which did not display strong growth rate changes, coalescence times aligned closely with neutral evolution predictions. Its phylogenetic tree was more balanced compared to the adapting cell lines, as shown in Figure 4.

These results demonstrate that phylogenetic trees constructed from single snapshots of transcriptomic data stored with CLONEID, can predict selection and growth rate changes in evolving cancer cell lines. These analyses highlight the potential of CLONEID to drive hypothesis-generating research.

Use Case 3: Inferring karyotype fitness landscapes from patient data

Type of Use Case: Mechanistic, Retrospective

Accumulating evidence highlights that the phenotypic impact of copy number alterations (CNAs) is highly context-dependent, with their effects influenced significantly by other concurrent genomic events. For instance, an analysis of chromosome-arm aneuploidies across 6,977 tumors identified numerous pairs of synergistically co-occurring aneuploidies that served as prognostic biomarkers, even though individual alterations lacked predictive power when considered independently of their broader karyotypic context [21]. We used CLONEID to record MRI imaging (Fig. 5A) and karyotype data (Fig. 5B) from a cohort of 11 glioblastoma (GBM) patients with matched primary and recurrent tumor specimens [22, 23]. We then applied ALFA-K to model the karyotype fitness landscape for each patient [24]. ALFA-K (Adaptive Local Fitness Landscapes for Aneuploid Karyotypes) is a computational framework designed to predict the fitness consequences of copy number alterations (CNAs) in a karyotype context-dependent manner, enabling prediction of tumor evolution from single-cell karyotype data. For instance, in patient P30, ALFA-K inferred the fitness of 28,405 unique karyotypes (Fig. 5C) based on the dynamics of 36 karyotypes observed in the input data (Fig. 5B). Cross-validation of the inferred fitness landscapes demonstrated strong concordance for 9 out of 11 cases, which were subsequently included in downstream analyses.

Across all nine landscapes, we identified 203,064 unique karyotypes (Fig. 5E), of which 9.47% (19,230 karyotypes) occurred in more than one landscape (hereafter referred to as shared karyotypes) and 0.02% (45 karyotypes) were present in all landscapes (ubiquitous karyotypes). We calculated the Pearson correlation coefficient between the fitness values of shared karyotypes across patient pairs, identifying significant positive and negative correlations in 13 and 20 out of 36 pairs, respectively (Fig. 5G; bubble chart above diagonal). Pearson correlation coefficient between the fitness values of ubiquitous karyotypes were higher (Fig. 5F), with 6 positive and 3 negative correlations (Fig. 5G; bubble chart below diagonal).

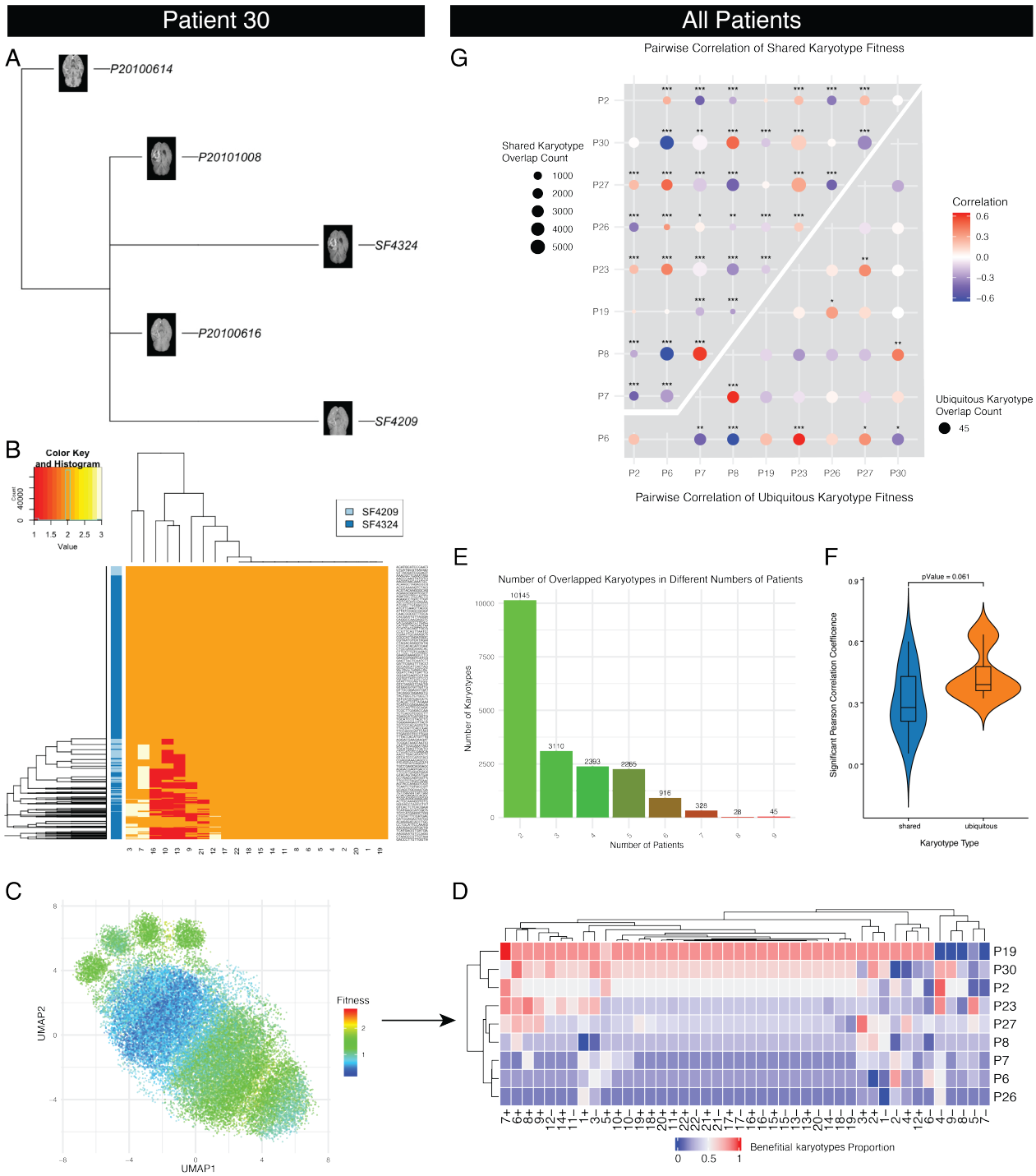

Figure 5: Multimodal analysis of a Glioblastoma patient through CLONEID informs a mathematical model of karyotype evolution. (A) MRI data from three timepoints were viewed through CLONEID's phenotype module; biopsies were taken during surgery at two timpoints (prefix SF*). (B) Karyotype composition derived from scRNA-seq of the two biopsies. 4,471 cells (rows) were sequenced in total and were used as input to a mathematical model for inferring karyotype fitness landscapes. (C) UMAP representation of 28,405 unique karyotypes that make up the charted fitness landscape for P30. Each karyotype is color-coded by its inferred fitness. (E) The histogram shows the karyotype count (Y-axis) shared across patient samples (X-axis). (F)Violin and box plots display the distribution of Pearson correlation coefficients between fitness and shared or ubiquitous karyotype groups (t-test). (G) Pairwise correlation coefficients between karyotypes shared by the given pair of patients (above diagonale) and between karyotypes shared by all patients (below diagonale). (D) Fitness consequence of copy number alterations (CNAs; x-axis) is shown for P30 along with all analyzed patients (y-axis). Entries calculated as proportion of karyotypes in C), for which a given CNA increases fitness.

This context-dependent fitness highlights the importance of further investigation into personalized factors driving variability in selection pressures among patients.

To further characterize karyotype selection in this GBM cohort, we computed the proportion of karyotypes in which a given copy number alteration (CNA) conferred a net fitness advantage (Fig. 5D). This analysis revealed that most CNAs are deleterious, while beneficial karyotypes are largely context-dependent. Overall, these findings demonstrate that karyotype fitness landscapes exhibit both patient-specific and shared selection pressures, underscoring the heterogeneous yet partially predictable nature of karyotype evolution in GBM.

4 Discussion

We present a framework for routine and systematic tracking of the pedigree of evolving cell populations over decades along with potentially changing environments (e.g. cell culture habits, therapy). Such efforts could help reveal long-term trends in the evolution of cell lines and tumors, that remain elusive at shorter time-scales. A system that tracks clones and how they change over time widens the path for genotype-phenotype mapping and discovery at large scale: e.g. in what direction do changes in culture conditions, or administration of specific therapeutics steer a population's subclone composition?

CLONEID is designed as a collaborative platform that bridges the gap between cancer biology and computational modeling, enabling seamless interaction between experimentalists and computational or mathematical biologists. Experiments to inform mathematical models are carefully planned to ensure they will provide the necessary information for modeling, a process that requires coordination and discussions between experimentalists and computational/mathematical biologists and often takes weeks per experiment. CLONEID provides a workflow to streamline this process. It ensures that the information needed for modeling is permanently stored in a centralized location with minimal extra burden on the experimentalist. For example, in addition to a cell-by-feature matrix, the lineage tracing module records time at which the underlying image was taken – a critical record for modeling that is often missing from routine experiments.

Consistency in the usage of CLONEID is key and crystallizes into several use cases: 1) Longitudinal information recorded by CLONEID informs parameters of mathematical models. This can be done with the lineage tracing module, with the multi-omics module or with both in combination depending on the hypothesis being tested with the mathematical model. 2) Genetic and morphologic features collected by the multi-omics and lineage tracing module respectively can be used for association studies investigating how genes affect cell morphology. 3) Long-term cell culture experiments where cells are challenged to adapt to a new environment (e.g. limiting Glucose availability). For such experiments, it is difficult to predict how long they will last (often several months or even years) and at what interval to sample cells for sequencing. Monitoring subtle morphological changes, such as changes in cell size, can inform experimentalists when to sample cells for sequencing. CLONEID facilitates this by automatically storing morphological features in the context of other information such as cell density, time, pedigree. 4) Routine recording of cell images over years opens the stage for deep learning applications, which require data at a scale beyond the scope of most experimental designs. Since its inception, CLONEID has stored over 30,000 microscopy images, each paired with rich metadata and embedded in the context of large pedigree trees (Figure 2A). Notably, this remarkable dataset has been generated by a single lab in just a few years. Scaling this platform to support research across entire laboratories and careers worldwide could unlock unparalleled insights and generate invaluable data for the broader scientific community.

While the primary goal of CLONEID is to provide infrastructure for efficient collaborations between experimentalists, computational biologists, and mathematicians, its design also offers a powerful means to quantify biological variability, which is a critical step toward improving reproducibility in cancer biology. More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own [25]. In oncology, 97% of clinical trials fail despite

promising preclinical results, with poor authentication of biological reagents being the leading cause of irreproducibility in in-vitro and in-vivo studies. While the precise contribution of cell line heterogeneity to this problem is unknown, there is substantial evidence that it plays a significant role. For example, work by Uri Ben-David et al. (2018) demonstrated that even slight differences in media composition can drive the selection of genetically distinct subpopulations, which in turn dramatically affect drug response. CLONEID addresses this challenge not just by controlling for variability, but by making it measurable and transparent. It systematically tracks (i) the pedigree of cell lineages, (ii) the exact media composition, and (iii) the number of cell divisions—an important proxy for the evolutionary time each clone experiences. By capturing this level of granularity, CLONEID enables researchers to interpret findings in the context of known variability.

We expect CLONEID to streamline acquisition of phenotypic measurements at high temporal resolution, as a byproduct of any *in-vitro/in-vivo* experiment, without complicating experimental design. Accumulating evidence that cancer cell lines are genetically heterogeneous and dynamic [16, 12, 13, 26], is primarily seen as a tremendous inconvenience. But this paradigm shift places these model systems closer to the clinical reality of primary tumors, which also adapt and change in response to their environment. Success in the clinical setting requires an experimental setting in which our therapeutic solutions are challenged with the adaptive nature of cancer. We seek to contribute to a paradigm shift in the way research laboratories define success of pre-clinical studies, not as the maximum reduction in cancer cell population growth, but as the ability to control a population's clonal dynamics over several generations of environmental adaptation.

Acknowledgments

We thank Sue Grimes for her early contributions to the database design, which helped ensure its long-term scalability.

Funding

This work was supported by the NCI grants 1R03CA259873-01A1, 1R37CA266727-01A1 and 1R21CA269415-01A1 awarded to N. Andor. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

References

[1] M. Greaves and C. C. Maley. Clonal evolution in cancer. Nature, 481(7381):306–313, January 2012.

[2] J. P. O'Connor, C. J. Rose, J. C. Waterton, R. A. Carano, G. J. Parker, and A. Jackson. Imaging Intratumor Heterogeneity: Role in Therapy Response, Resistance, and Clinical Outcome. Clinical cancer research : an official journal of the American Association for Cancer Research, 21(2):249–257, January 2015. PMID: 25421725.

[3] A. Kreso, C. A. O'Brien, P. van Galen, O. I. Gan, F. Notta, A. M. K. Brown, K. Ng, J. Ma, E. Wienholds, C. Dunant, A. Pollett, S. Gallinger, J. McPherson, C. G. Mullighan, D. Shibata, and J. E. Dick. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science (New York, N.Y.), 339(6119):543–548, February 2013. PMID: 23239622.

[4] C. Ma, M. Balaban, J. Liu, S. Chen, M. J. Wilson, C. H. Sun, L. Ding, and B. J. Raphael. Inferring allele-specific copy number aberrations and tumor phylogeography from spatially resolved transcriptomics. Nature Methods, 21(12):2239–2247, December 2024.

[5] S. Zaccaria and B. J. Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. Nature Communications, 11(1):4301, September 2020.

[6] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: statistical inference of clonal population structure in cancer. Nature Methods, 11(4):396–398, April 2014.

[7] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (New York, N.Y.), 344(6190):1396–1401, June 2014. PMID: 24925914.

[8] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. Nature, 472(7341):90–94, April 2011.

[9] H. S. Kim, S. M. Grimes, T. Chen, A. Sathe, B. T. Lau, G.-H. Hwang, S. Bae, and H. P. Ji. Direct measurement of engineered cancer mutations and their transcriptional phenotypes in single cells. Nature Biotechnology, 42(8):1254–1262, August 2024.

[10] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. Nature Methods, 18(1):100–106, January 2021.

[11] L. K. Shiyam Sundar, J. Yu, O. Muzik, O. C. Kulterer, B. Fueger, D. Kifjak, T. Nakuz, H. M. Shin, A. K. Sima, D. Kitzmantl, R. D. Badawi, L. Nardo, S. R. Cherry, B. A. Spencer, M. Hacker, and T. Beyer. Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine, 63(12):1941–1948, December 2022. PMID: 35772962.

[12] U. Ben-David, B. Siranosian, G. Ha, H. Tang, Y. Oren, K. Hinohara, C. A. Strathdee, J. Dempster, N. J. Lyons, R. Burns, A. Nag, G. Kugener, B. Cimini, P. Tsvetkov, Y. E. Maruvka, R. O'Rourke, A. Garrity, A. A. Tubelli, P. Bandopadhayay, A. Tsherniak, F. Vazquez, B. Wong, C. Birger, M. Ghandi, A. R. Thorner, J. A. Bittker, M. Meyerson, G. Getz, R. Beroukhim, and T. R. Golub. Genetic and transcriptional evolution alters cancer cell line drug response. Nature, 560(7718):325–330, August 2018. PMID: 30089904.

[13] U. Ben-David, G. Ha, Y.-Y. Tseng, N. F. Greenwald, C. Oh, J. Shih, J. M. McFarland, B. Wong, J. S. Boehm, R. Beroukhim, and T. R. Golub. Patient-derived xenografts undergo mouse-specific tumor evolution. Nature Genetics, 49(11):1567–1575, November 2017. PMID: 28991255.

[14] T. Li, J. Liu, J. Feng, Z. Liu, S. Liu, M. Zhang, Y. Zhang, Y. Hou, D. Wu, C. Li, Y. Chen, H. Chen, and X. Lu. Variation in the life history strategy underlies functional diversity of tumors. National Science Review, 8(nwaa124), February 2021.

[15] P. Hughes, D. Marshall, Y. Reid, H. Parkes, and C. Gelber. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? BioTechniques, 43(5):575, 577–578, 581–582 passim, November 2007. PMID: 18072586.

[16] N. Andor, B. T. Lau, C. Catalanotti, A. Sathe, M. A. Kubit, J. Chen, S. M. Grimes, C. J. Suarez, and H. P. Ji. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. NAR Genomics and Bioinformatics (in press), March 2020.

[17] Y. Liu, Y. Mi, T. Mueller, S. Kreibich, E. G. Williams, A. Van Drogen, C. Borel, M. Frank, P.-L. Germain, I. Bludau, M. Mehnert, M. Seifert, M. Emmenlauer, I. Sorg, F. Bezrukov, F. S. Bena, H. Zhou, C. Dehio, G. Testa, J. Saez-Rodriguez, S. E. Antonarakis, W.-D. Hardt, and R. Aebersold. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nature Biotechnology, 37(3):314–322, March 2019.

[18] S. H. Selznick, M. L. Thatcher, K. S. Brown, and C. A. Haussler. Development and application of computer software for cell culture laboratory management. In Vitro Cellular & Developmental Biology. Animal, 37(1):55–61, January 2001. PMID: 11249207.

[19] S. G. Canfield, G. Jin, S. P. Palecek, and T. Sampsell. Software to improve transfer and reproducibility of cell culture methods. BioTechniques, 65(5):289–292, November 2018.

[20] S. Müller, A. Cho, S. J. Liu, D. A. Lim, and A. Diaz. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. Bioinformatics (Oxford, England), 34(18):3217–3219, September 2018. PMID: 29897414.

[21] A. Shukla, T. H. M. Nguyen, S. B. Moka, J. J. Ellis, J. P. Grady, H. Oey, A. S. Cristino, K. K. Khanna, D. P. Kroese, L. Krause, E. Dray, J. L. Fink, and P. H. G. Duijf. Chromosome arm aneuploidies shape tumour evolution and drug response. Nature Communications, 11(1):449, January 2020.

[22] S. Müller, E. D. Lullo, A. Bhaduri, B. Alvarado, G. Yagnik, G. Kohanbash, M. Aghi, and A. Diaz. A single-cell atlas of human glioblastoma reveals a single axis of phenotype in tumor-propagating cells. bioRxiv, page 377606, July 2018.

[23] L. Wang, J. Jung, H. Babikir, K. Shamardani, S. Jain, X. Feng, N. Gupta, S. Rosi, S. Chang, D. Raleigh, D. Solomon, J. J. Phillips, and A. A. Diaz. A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets. Nature Cancer, 3(12):1534–1552, December 2022.

[24] R. J. Beck and N. Andor. Local Adaptive Mapping of Karyotype Fitness Landscapes, July 2023. Pages: 2023.07.14.549079 Section: New Results.

[25] M. Baker. 1,500 scientists lift the lid on reproducibility. Nature, 533(7604):452–454, May 2016. PMID: 27225100.

[26] D. C. Minussi, M. D. Nicholson, H. Ye, A. Davis, K. Wang, T. Baker, M. Tarabichi, E. Sei, H. Du, M. Rabbani, C. Peng, M. Hu, S. Bai, Y.-W. Lin, A. Schalck, A. Multani, J. Ma, T. O. McDonald, A. Casasent, A. Barrera, H. Chen, B. Lim, B. Arun, F. Meric-Bernstam, P. Van Loo, F. Michor, and N. E. Navin. Breast tumours maintain a reservoir of subclonal diversity during expansion. Nature, 592(7853):302–308, April 2021. PMID: 33762732.